Revue des sciences de l'éducation

L'interprétation des tests d'hypothèses : *p,* la taille de l'effet et la puissance

Interpreting null-hypothesis significance tests : *p,* effect size, and power

La interpretación de las pruebas de hipótesis : *p,* el tamaño del efecto y la potencia

Jimmy Bourque, Jean-Guy Blais et François Larose

Volume 35, numéro 1, 2009

Avoir des difficultés scolaires importantes à l'école : quelles formules, quel avenir ?

URI : https://id.erudit.org/iderudit/029931ar
DOI : https://doi.org/10.7202/029931ar

Aller au sommaire du numéro

Éditeur(s)

Revue des sciences de l'éducation

ISSN

0318-479X (imprimé)
1705-0065 (numérique)

Découvrir la revue

Citer cet article

Bourque, J., Blais, J.-G. & Larose, F. (2009). L'interprétation des tests d'hypothèses : *p,* la taille de l'effet et la puissance. *Revue des sciences de l'éducation*, *35*(1), 211–226. https://doi.org/10.7202/029931ar

Résumé de l'article

Cet article vise à expliciter la logique des tests d'hypothèses en recherche. Plusieurs études ont démontré une confusion quant à la signification des résultats de tests d'hypothèses. Cette confusion proviendrait en partie des points divergents entre les approches de Fisher, de Neyman et Pearson ainsi que de Bayes. L'article précise l'information fournie par le coefficient de signification, la taille de l'effet et la puissance et propose différents logiciels permettant l'analyse de la puissance et de la taille de l'effet.

L'interprétation des tests d'hypothèses : *p*, la taille de l'effet et la puissance

Jimmy Bourque, professeur
Université de Moncton

Jean-Guy Blais, professeur
Université de Montréal

François Larose, professeur
Université de Sherbrooke

RÉSUMÉ • Cet article vise à expliciter la logique des tests d'hypothèses en recherche. Plusieurs études ont démontré une confusion quant à la signification des résultats de tests d'hypothèses. Cette confusion proviendrait en partie des points divergents entre les approches de Fisher, de Neyman et Pearson ainsi que de Bayes. L'article précise l'information fournie par le coefficient de signification, la taille de l'effet et la puissance et propose différents logiciels permettant l'analyse de la puissance et de la taille de l'effet.

MOTS CLÉS • tests d'hypothèses, signification statistique, taille de l'effet, puissance, inférence.

1. Introduction

Lorsque des chercheurs lisent ou publient des résultats de tests d'hypothèses, ils portent généralement une attention particulière aux coefficients de signification (*p*). Dans le second cas, ils espèrent que leur valeur sera inférieure ou égale à 0,05. De plus, la plupart des rédacteurs et évaluateurs de revues savantes ont les mêmes préoccupations, d'où, peut-être, une partie de leur intérêt pour *p*. Le fait d'obtenir des résultats statistiquement significatifs accroît leurs chances d'être publiés (Maddock et Rossi, 2001 ; Nakagawa, 2004 ; Poitevineau, 2004).

Or, selon plusieurs auteurs, la valeur de *p* ne devrait constituer qu'une étape de l'interprétation des tests d'hypothèses. En fait, il faudrait, généralement, que son rôle soit appuyé par d'autres informations comme la taille de l'effet ; mais, dans tous les cas, il n'aurait qu'une importance pratique limitée (Cohen, 1962 ; Gigerenzer, 1993 ; Kline, 2004 ; Thompson, 1989).

2. Problématique

Depuis le début des années 1960, des études continuent d'identifier un recours presque exclusif à la valeur de *p* pour l'interprétation des tests d'hypothèses et

affichent, de plus, une puissance expérimentale insuffisante (Bezeau et Graves, 2001 ; Cashen et Geiger, 2004 ; Clark-Carter, 1997 ; Cohen, 1962 ; Jennions et Møller, 2003 ; Kosciulek et Szymanski, 1993 ; Maddock et Rossi, 2001 ; Mone, Mueller et Mauland, 1996 ; Paul et Plucker, 2004 ; Vacha-Haase et Nilsson, 1998). Cette situation persiste même si de nombreux auteurs ont tenté de provoquer un changement dans les pratiques en sciences humaines et sociales (Carver, 1978 ; Cohen, 1994 ; Meehl, 1978 ; Morrison et Henkel, 1970 ; Shrout, 1997 ; Thomas et Juanes, 1996). Les préoccupations quant aux répercussions de cette situation sur la rigueur de la recherche sont assez importantes pour que soit créé, aux États-Unis, un groupe de travail sur l'inférence statistique (Wilkinson and the Task Force on Statistical Inference, 1999). Des réactions similaires proviennent également du monde francophone (Giguère, Hélie et Cousineau, 2004 ; Lecoutre, 1982 ; Lecoutre et Poitevineau, 2000 ; Lecoutre, Poitevineau et Lecoutre, 2005 ; Michel, Ollivier-Gay, Spiegel et Boutin, 2002 ; Poitevineau, 2004 ; Rouanet, 1991). Au Québec, en 1991, Blais proposait une réflexion sur la pratique de la statistique en éducation. Ce texte s'avère d'ailleurs un précurseur du présent article mais, comme il a été publié dans une revue avec tirage limité et maintenant disparue, il peut être difficilement accessible à la communauté scientifique. Par conséquent, il paraît approprié de réactualiser et de compléter le propos puisque, outre ce texte, la préoccupation envers l'adéquation de l'usage des tests d'hypothèses dans la francophonie ne semble pas déborder le cadre de la psychologie et de la médecine.

3. Objectif de recherche

Dans cet article, nous souhaitons faire le point sur l'utilisation des tests d'hypothèses sur les aléas du recours exclusif à la valeur de *p*. Plus précisément, après avoir présenté un bref historique des tests d'hypothèses, nous aborderons les concepts de signification statistique, de taille de l'effet et de puissance des tests statistiques ; ensuite, nous présenterons quelques outils permettant l'analyse de la taille de l'effet et de la puissance statistique lors de la réalisation de tests d'hypothèses.

4. Cadre théorique : les tests d'hypothèses

Le test d'hypothèses ne trouve son utilité que lorsque l'étude de la population entière est impossible et que le chercheur doit plutôt analyser un échantillon de cette population (Blais, 1991). Dans ce cas, comme l'échantillonnage comporte inévitablement une marge d'erreur, le test d'hypothèses vise à indiquer la probabilité d'obtenir les statistiques observées sur la base d'une hypothèse quant à la valeur d'un paramètre de la population. Si le premier test d'hypothèses connu, le test du khi-carré, peut être attribué à Karl Pearson (1857-1936), c'est Ronald Fisher (1890-1962) qui a d'abord esquissé la logique méthodologique des tests d'hypothèses.

4.1 L'approche de Fisher

Selon Fisher, dans le cadre d'un plan expérimental, le test d'hypothèses vise à réfuter une hypothèse donnée, sans lui adjoindre d'hypothèse concurrente. La logique de Fisher débute donc avec la formulation d'une hypothèse H, selon laquelle la statistique (la moyenne, par exemple) d'un échantillon aléatoire, tiré d'une population hypothétique infinie, est égale à une valeur donnée. Ensuite, on teste la différence entre le paramètre de la distribution d'échantillonnage théorique et la statistique observée dans l'échantillon. L'hypothèse sera rejetée si les valeurs comparées diffèrent de plus d'un écart convenu d'avance (Blais, 1991 ; Chow, 1996). Au départ, une certaine confusion peut provenir de l'évolution dans le temps du discours de Fisher : dans ses premiers écrits, il prône l'adoption de critères de signification fixes, alors que, dans les années 1950, il change de position et propose que ces critères puissent être variables. Le chercheur devrait alors publier la probabilité exacte obtenue, et non la valeur retenue du critère de signification (Gigerenzer, 1993). Par ailleurs, en l'absence de résultats *significatifs*, c'est-à-dire de résultats permettant de rejeter H, l'hypothèse n'est pas acceptée : le chercheur suspend alors son jugement. Selon Fisher, la finalité des tests était l'inférence inductive, bien que la probabilité obtenue soit d'obtenir les données observées en postulant la véracité de l'hypothèse nulle, donc $P(D|H)$.

4.2 L'approche de Neyman et Pearson

La contribution de Jerzy Neyman (1894-1981) et d'Egon Pearson (1895-1980) se voulait une tentative de consolider les travaux de Fisher de la transformer en une approche plus cohérente et rigoureuse (Gigerenzer, 1993). Par conséquent, Neyman et Pearson délaissent l'inférence inductive pour mettre les tests d'hypothèses au service de la prise de décision dans des contextes pragmatiques. Ainsi, ils ajoutent notamment à l'approche de Fisher une analyse dans une logique de coûts et de bénéfices (Chow, 1996). D'abord, là où Fisher ne posait qu'une seule hypothèse, Neyman et Pearson formulent une hypothèse testée (H_0) et une contre-hypothèse ou hypothèse alternative (H_1). Ces deux hypothèses se doivent d'être exhaustives et mutuellement exclusives, de sorte que le rejet de l'une implique l'acceptation de l'autre, et vice-versa (Poitevineau, 2004). Il s'ensuit l'introduction de deux types d'erreurs et de leur probabilité associée : l'erreur de type I (α), soit rejeter H_0 à tort, et l'erreur de type II (β), soit conserver H_0 à tort. Dans cette logique, les valeurs acceptables de α et β sont fixées *a priori* : il appartient au chercheur de déterminer les risques d'erreurs qu'il est prêt à assumer, et ce, en tenant compte des coûts relatifs à chaque type d'erreurs. Soulignons qu'avec le concept d'erreur de type II apparaît aussi le concept de puissance statistique, qui est son complément en termes de probabilités ($1 - \beta$). De plus, c'est le caractère pragmatique de l'approche de Neyman et Pearson qui confère son utilité au calcul de la taille des effets observés (Blais, 1991). Précisons finalement que, contrairement à Fisher, qui posait l'hypothèse d'une population infinie pour pouvoir utiliser le concept de distribution

d'échantillonnage, Neyman et Pearson entrevoient leur interprétation des tests d'hypothèses dans un contexte de répétition ; ainsi, la probabilité α devient le pourcentage d'erreurs de type I, commises par le chercheur sur une grande série de répétitions de la même expérience et provenant du tirage successif d'échantillons aléatoires d'une même population (Blais, 1991).

4.3 L'approche de Bayes

Les deux approches discutées précédemment appartiennent toutes deux à l'allégeance *fréquentiste*, pour emprunter le terme anglo-saxon, par opposition à l'approche bayésienne. Cette dernière découle des travaux de Thomas Bayes (1702-1761) et se distingue des approches précédentes en ce qu'elle vise à établir le degré de certitude par rapport à une hypothèse sur la base des données obtenues. Il s'agit donc de P(H|D), soit la probabilité de l'hypothèse H conditionnelle à l'observation des données D, le Saint Graal de l'inférence statistique, à une nuance près, c'est-à-dire le degré de *certitude* envers la vraisemblance de l'hypothèse et non de la probabilité que l'hypothèse soit effectivement vraie. Or, le calcul de cette vraisemblance (évaluée comme une probabilité) requiert notamment du chercheur qu'il attribue une probabilité initiale à la véracité de l'hypothèse, ce qui devient subjectif dans la mesure où cette probabilité *théorique* est généralement inconnue. Ce procédé devient d'autant plus subjectif que l'ignorance du chercheur quant au phénomène observé est grande. De plus, comme plusieurs chercheurs pourraient attribuer des probabilités initiales différentes à une même hypothèse, le procédé risque de refléter davantage les opinions du chercheur que la réalité (Blais, 1991 ; Chow, 1996).

5. En pratique…

En pratique, que ce soit dans les manuels de statistique, dans le cadre de la formation des chercheurs ou dans les textes que publient les revues savantes en sciences humaines et sociales (Clark-Carter, 1997 ; Giguère, Hélie et Cousineau, 2004 ; Poitevineau, 2004), l'interprétation des tests d'hypothèses fait appel à une logique mixte combinant des éléments des trois approches présentées. D'abord, en accord avec Fisher, seule l'hypothèse nulle est généralement formulée (lorsqu'elle l'est explicitement), la visée est surtout l'inférence inductive où, conformément à sa position la plus récente, plusieurs seuils de signification sont utilisés (0,05 ; 0,01 et 0,001) et, souvent, les probabilités exactes sont publiées. Puis, dans la veine des travaux de Neyman et Pearson, apparaissent les concepts d'erreurs de types I et II et, plus rarement, des considérations sur la taille de l'effet (maintenant largement utilisée en psychologie) et la puissance statistique. Enfin, l'interprétation des résultats est souvent bayésienne, dans le sens où l'on généralise régulièrement les résultats des tests d'hypothèses en les associant à la population de référence (et ce, bien que la probabilité conditionnelle obtenue en réalité aille dans l'autre sens).

5.1 La signification statistique

Essayons maintenant d'illustrer, par un exemple, l'application de cette logique hybride des tests d'hypothèses. Supposons que nous voulons tester la différence entre garçons et filles d'un cours de sciences quant à la fréquence à laquelle ils ont à recopier les notes que leur enseignant écrit au tableau (Conseil des ministres de l'Éducation du Canada, 2007). Notons que, comme garçons et filles fréquentent les mêmes classes, il n'y a pas lieu de croire en la présence d'une différence significative. L'approche fréquentiste attribue une probabilité à la valeur d'une statistique obtenue par calcul à partir des données de recherche et testée à partir d'une distribution de probabilités connue (Blais, 1991 ; Chow, 1996 ; Loftus, 1996). C'est le cas, par exemple, du test du khi-carré (χ^2) : la différence entre les fréquences attendues et les fréquences observées permet de calculer la valeur de la statistique χ^2, à laquelle est attribuée une probabilité p sur la base de la distribution de probabilités du khi-carré. Dans notre exemple, nous cherchons à explorer l'association entre le sexe de l'élève et la pratique de l'enseignant, exprimée comme une fréquence parmi quatre catégories possibles. Notre échantillon compte 21 961 élèves et nous obtenons un résultat *statistiquement significatif* ($\chi^2[6] = 29{,}00$, $p = 0{,}00$) au seuil de 0,05. Ici, comme il est illogique de croire que garçons et filles ne sont pas exposés à la même quantité de cours magistraux, que signifie ce résultat *significatif* ? C'est que le résultat du test d'hypothèses sera déclaré *statistiquement significatif* ou non sur la base de la valeur de *p*. Ainsi, si *p* s'avère inférieur à un seuil de signification arbitraire, généralement 0,05 en sciences sociales, le résultat est dit statistiquement significatif. Or, sous l'influence de Fisher et de ses derniers écrits, la valeur de *p* est souvent mise en parallèle avec plusieurs valeurs critères, chaque seuil semblant correspondre à une certitude plus grande quant au rejet de l'hypothèse nulle, ou alors, la valeur exacte de *p* sera communiquée ($p = 0{,}00$ dans notre exemple) et considérée significative ou non à partir des mêmes critères. Ici, *p* (0,00) désigne la probabilité d'obtenir une valeur de χ^2 supérieure ou égale à 29,00 si cette valeur est de 0 dans la population (l'hypothèse nulle postule un effet d'une ampleur donnée, généralement nulle, dans la population hypothétique d'où serait tiré l'échantillon). Le rejet de cette hypothèse nulle (et de sa conséquence statistique, $\chi^2 = 0$) constitue d'ailleurs un abus de langage puisqu'elle est, pratiquement, toujours fausse (Kline, 2004). Il faut ici comprendre que, comme l'explique Chow (1996), il est théoriquement possible que l'hypothèse nulle soit vraie : étant donné des échantillons aléatoires tirés d'une population théorique, poser l'hypothèse nulle revient à soutenir qu'il n'y a, dans les données, aucune autre variation que celle attribuable aux fluctuations aléatoires d'échantillonnage. Or, cette condition correspond au *système isolé sans frottement* des physiciens : il s'agit d'une condition idéale, théorique, rarement rencontrée dans la pratique, puisqu'un nombre important de variables confondues, dont l'effet peut être plus ou moins important, viennent souvent menacer le postulat selon lequel les groupes comparés sont équivalents au départ. De plus, le chercheur ne dispose pratiquement jamais

d'échantillons aléatoires (Kline, 2004), ne serait-ce qu'à cause de l'obligation éthique d'obtenir le consentement des sujets qui donne un caractère volontaire à presque tout échantillon, qu'il soit originellement aléatoire ou non. Justement pour ces raisons, l'ouvrage de Chow (1996) a d'ailleurs suscité un vif débat, dans la revue *Behavioral and Brain Sciences*, quant à l'applicabilité pratique de son raisonnement théorique qui, lui, n'est pas contesté. Dans ce numéro, dirigé par Paul Bloom (Yale University, USA) et Barbara L. Finlay (Cornell University, USA), trente-sept auteurs (dont Gigerenzer, Poitevineau, Rouanet, Rossi et Cohen) ont réagi à l'ouvrage de Chow en publiant de courts articles. Nous avons repris les arguments énumérés dans ces articles, mais à partir d'autres textes, plus étoffés, dont certains provenant des auteurs ci-haut mentionnés.

Notons ici que les tests d'hypothèses postulent que l'hypothèse nulle est exactement vraie dans la population, pour ensuite calculer une probabilité d'obtenir les résultats observés à partir d'un échantillon aléatoire tiré de cette population. La probabilité conditionnelle évaluée est donc $P(D|H_0)$ et non $P(H_0|D)$ (Chow, 1996). Dans ce cas-ci, il s'agit de la probabilité d'observer *D* conditionnellement à la véracité de l'hypothèse nulle. Contrairement à une conception erronée, mais néanmoins largement répandue, les résultats de tests d'hypothèses ne donnent pas d'informations sur la population dans la mesure où la probabilité que l'hypothèse nulle y soit exactement vraie est, pratiquement, infinitésimale (Jones et Tukey, 2000 ; Loftus, 1996 ; Millis, 2003, Vacha-Haase et Thompson, 1998). Par contre, *une valeur de p [statistiquement] non significative ne signifie pas qu'il n'y a pas de différence mais plutôt qu'aucune preuve de l'existence d'une différence n'a pu être trouvée* (Millis 2003, p. 222, traduction libre).

5.2 La taille de l'effet

Dans une optique pragmatique, l'une des conceptions associées à la signification statistique est que l'on confond *signification* et *importance*. Ainsi, un résultat statistiquement significatif ne se traduit pas nécessairement par une importance au niveau pratique. Il faut ici considérer qu'un échantillon suffisamment vaste mènera généralement au rejet de l'hypothèse nulle ou, en d'autres mots, à un résultat statistiquement significatif et ce, peu importe la taille réelle de l'effet (pourvu qu'elle ne soit pas nulle).

La taille de l'effet désigne *à quel degré un phénomène donné est présent dans la population* (Cohen 1988, p. 9, traduction libre). Ainsi, une autre façon de concevoir l'hypothèse nulle est de dire que la taille de l'effet est nulle. Dans cette optique, la taille de l'effet décrit dans quelle mesure l'hypothèse nulle est fausse : plus la taille de l'effet est grande, plus il est justifié de rejeter l'hypothèse nulle. Or, pour un grand échantillon, un effet même minime suffira à la faire rejeter. Concrètement, ce problème peut être envisagé à travers l'exemple présenté plus haut : notre résultat significatif au test du khi-carré s'accompagne d'un effet (calculé à l'aide du *V* de Cramer) considéré comme significatif, mais néanmoins négligeable, selon les

critères de Cohen ($V = 0{,}02$; $p = 0{,}00$). La portée pratique d'une telle conclusion peut difficilement être vue comme importante. Néanmoins, comme lors de l'interprétation de *p*, il faut éviter de porter un jugement mécanique sur la taille de l'effet (Sawilowsky, 2003). Dans la mesure où les exigences du test d'hypothèses sont respectées, un résultat statistiquement non significatif malgré une taille d'effet importante signifie que cet effet, aussi grand soit-il, pourrait être obtenu avec une probabilité supérieure à la valeur critère (Chow, 1996). Pour éviter une telle situation, le chercheur doit veiller, notamment, à minimiser l'erreur de mesure. Cela dit, afin de fournir au lecteur toute l'information requise pour juger des résultats obtenus, le chercheur devrait toujours fournir la taille de l'effet au terme d'un test d'hypothèse (American Psychological Association, 2001 ; Wilkinson and the Task Force on Statistical Inference, 1999). En fait, la taille de l'effet peut s'avérer plus utile à l'interprétation des résultats que la valeur de *p*, surtout si l'ampleur de cet effet peut influencer les décisions pratiques subséquentes (Kline, 2004).

5.3 La puissance statistique

Supposons maintenant qu'une biologiste place une lamelle sous son microscope pour vérifier la présence de bactéries et qu'elle constate que la lamelle semble vierge. Peut-elle en déduire que c'est obligatoirement le cas ? En fait, deux explications sont possibles : 1) la lamelle est effectivement exempte de bactéries ou 2) l'objectif n'est pas assez puissant pour distinguer des corps aussi petits que des bactéries. Notons que, dès qu'une bactérie est visible, la question ne se pose pas.

Dans la logique de Neyman et Pearson (le concept de puissance statistique ne s'applique pas à l'approche de Fisher), le même raisonnement peut être appliqué aux tests d'hypothèses. Lorsque le chercheur obtient un résultat non significatif, est-ce dû au fait que cet échantillon pourrait provenir d'une population où l'hypothèse nulle est vraie ou est-ce dû à un manque de puissance statistique ? Encore ici, la question n'a de sens qu'en l'absence de résultats statistiquement significatifs.

Selon Cohen (1988), la puissance d'un test statistique peut se définir comme *la probabilité qu'il produise des résultats statistiquement significatifs* (p. 1, traduction libre) quand H_0 est fausse. Au regard d'une hypothèse nulle donnée, la puissance peut aussi être comprise comme la probabilité qu'un test donné mène à son rejet. La puissance d'un test statistique dépend principalement de trois paramètres : le seuil de signification, la taille de l'échantillon et la taille de l'effet.

Ainsi, quand le seuil de signification est augmenté (par exemple, passe de 0,05 à 0,10), la puissance augmente aussi. La direction d'un test a également une influence sur la puissance. Un test statistique peut être bilatéral (*two-tailed*) ou unilatéral (*one-tailed*). Dans un test bilatéral, l'hypothèse nulle peut être rejetée sur la base d'un écart dans chaque direction par rapport à l'hypothèse nulle. Au contraire, dans un test unilatéral, un écart dans une seule direction permet de rejeter l'hypothèse nulle. Les tests unilatéraux sont plus puissants que les tests bilatéraux,

mais leur puissance est nulle dans la direction autre que celle spécifiée (Cohen, 1988 ; Kline, 2004).

Par ailleurs, la *fiabilité* d'un estimateur provenant d'un échantillon désigne la précision avec laquelle il parvient à identifier approximativement la valeur du paramètre correspondant relié à la population de référence. En d'autres termes, un estimateur fiable est celui qui minimise l'erreur due à l'échantillonnage. La taille de cette erreur varie de façon inverse au nombre d'unités statistiques incluses dans l'échantillon : plus l'échantillon est grand, plus l'erreur sera faible et plus l'estimateur sera fiable. Conséquemment, quand la taille de l'échantillon augmente, la puissance statistique augmente aussi. À la limite, comme la probabilité de retrouver un effet exactement nul dans la population est pratiquement inexistante, tout échantillon suffisamment vaste finira par produire des résultats significatifs (Vacha-Haase et Nilsson, 1998). Enfin, plus l'effet est grand, plus il sera facile à détecter. Par conséquent, si la taille de l'effet augmente, la puissance statistique augmente aussi.

En outre, comme le fait remarquer Maxwell (2004), de nombreux devis de recherche impliquent des tests d'hypothèses multiples, par exemple, lors de l'utilisation de régressions multiples ou d'analyses de variance à plan factoriel. Ainsi, une ANOVA donnée peut mener au calcul de trois puissances statistiques distinctes : la probabilité de détecter au moins un effet significatif, la probabilité de détecter un effet significatif donné et la probabilité de détecter tous les effets significatifs. Ces trois puissances statistiques seront généralement d'ampleurs différentes : la probabilité de détecter au moins un effet sera plus élevée que celle de détecter un effet donné, qui à son tour sera plus élevée que celle de détecter tous les effets. Conséquemment, une étude peut démontrer une puissance statistique suffisante pour détecter au moins un effet, mais rendre la détection de tous les effets peu probable. Autrement dit, il se peut que la probabilité de commettre une erreur de type II pour au moins un effet soit très élevée malgré une puissance statistique suffisante pour détecter au moins un effet (la puissance peut aussi être considérée comme la probabilité complémentaire à la probabilité d'erreur de type II : $1 - \beta$). Certains auteurs (Baguley, 2004 ; Biskin, 1998), à la suite de Neyman et Pearson, questionnent d'ailleurs la priorité accordée systématiquement à l'erreur de type I dans les écrits : selon la situation étudiée, l'erreur de type II peut avoir des conséquences pratiques plus graves que l'erreur de type I. Alors, pourquoi cette préoccupation à se prémunir uniquement contre cette dernière ? Les seuils acceptables pour α et β ne devraient-ils pas dépendre des coûts relatifs associés à chaque type d'erreurs dans la situation étudiée ?

6. Application

6.1 Analyse *a priori*

Il existe deux types d'analyse de puissance : l'analyse *a priori* et l'analyse *a posteriori*. L'analyse *a priori* a lieu avant le recueil des données et vise à établir le nombre

d'unités statistiques nécessaires à l'obtention de résultats significatifs moyennant un effet donné. Le chercheur choisit alors un seuil de signification, généralement 0,05, et une puissance statistique minimale pour la détection de l'effet souhaité. Il faut ensuite déterminer à partir de quelle taille le chercheur désire qu'un effet soit détecté. Il faut ici comprendre que tout effet n'est pas nécessairement intéressant. Ainsi, que la puissance soit insuffisante pour détecter un effet presque nul ne pose pas nécessairement problème. Pour fixer la taille de l'effet que le chercheur souhaite pouvoir détecter, une des façons de faire est de consulter les résultats d'autres recherches sur le phénomène étudié afin de déterminer la taille habituelle des effets observés. Cette information fournit au moins un ordre de grandeur au chercheur. Dans le cas où cette information ne serait pas disponible, des règles informelles ont été énoncées par Cohen (1988), qui divise les effets selon qu'ils sont de taille négligeable, petite, moyenne ou grande. Toutefois, comme toute règle de ce type, elles ne tiennent pas compte de la spécificité du phénomène étudié et constituent un dernier recours pour l'interprétation des résultats. À partir de ces informations (taille minimale de l'effet que l'on désire détecter, puissance minimale et seuil de signification), le chercheur pourra déterminer la taille minimale que devrait avoir son échantillon pour obtenir une probabilité fixée d'avance de détecter l'effet désiré au seuil de signification souhaité.

6.2 Analyse *a posteriori*

L'analyse *a posteriori* de la puissance statistique et, surtout, de la taille de l'effet, est effectuée après le recueil des données, et permet de connaître la puissance statistique des tests effectués et de mettre les résultats en perspective. Or, comme le fait justement remarquer Kline (2004), l'analyse de puissance *a posteriori* s'apparente à une autopsie : si la puissance s'avère insuffisante, il est trop tard pour prévenir ou guérir. La taille de l'effet s'avère alors particulièrement pertinente. Deux cas nécessitent tout spécialement une analyse plus détaillée. Dans le premier cas, le chercheur détecte un effet minime, mais significatif. L'analyse de puissance *a posteriori* révélera probablement une puissance élevée. C'est le cas de notre exemple : la probabilité de détecter l'effet négligeable obtenu à partir du test du khi-carré au seuil de 0,05 est de 80,35 %. Dans ce cas, la taille de l'effet peut aider à nuancer le jugement et à évaluer l'importance pratique du résultat obtenu : ici, en considérant la grande taille de l'échantillon et l'ampleur négligeable de l'effet observé, nous pourrions conclure que, bien que cet effet soit significatif, il demeure trop minime pour démontrer une importance pratique. Dans le second cas, le chercheur obtient un résultat non significatif et, habituellement, une puissance faible. Encore une fois, le chercheur doit interpréter la taille de l'effet pour déterminer si son résultat est dû à un effet négligeable (ce qui ne pose pas problème) ou à un autre facteur, comme un échantillon trop petit (donc une puissance trop faible) pour détecter un effet réel dont la taille mesurée pourrait revêtir une importance pratique (Loftus, 1996 ; Rosenthal, Rosnow et Rubin, 2000). Toutefois, notons que la

puissance *a posteriori* sera hautement liée à la valeur de *p* et ne peut pas être utilisée comme argument en faveur du maintien ou du rejet de l'hypothèse nulle (Baguley, 2004).

6.3 Quelques outils

Si, dans certains cas, la taille de l'effet et la puissance statistique peuvent être obtenues directement de logiciels statistiques polyvalents, il n'en est pas toujours ainsi. Pour des mesures d'association entre variables catégorielles (khi-carré), la taille de l'effet peut être interprétée à partir de phi (pour deux variables dichotomiques) ou du *V* de Cramer. Pour la corrélation de Pearson, la valeur du coefficient de corrélation *r* constitue un estimateur de la taille de l'effet, alors que R^2 peut être interprété en ce sens comme une régression linéaire. Cependant, le logiciel *SPSS* ne permet pas de calculer la puissance pour ces tests. Par ailleurs, les analyses de variance (ANOVA univariée et multivariée, ANOVA à mesures répétées) permettent le calcul, sur commande, de êta-carré (η^2), un estimateur de la taille de l'effet et de la puissance statistique. Comme il ne tient pas compte de la valeur de l'erreur, cet estimateur s'avère toutefois biaisé dans la mesure où il surestime systématiquement la taille réelle de l'effet. Des estimateurs non biaisés, tel oméga-carré (ω^2), devraient lui être préférés. Pour interpréter ces indices, Cohen (1988) fournit des critères d'interprétation de la taille de l'effet qui, s'ils ne constituent pas une solution idéale, vu leur caractère général et arbitraire, donnent toutefois un ordre de grandeur pour les effets observés.

Pour obtenir un estimateur de la taille de l'effet et la puissance statistique d'autres tests, le chercheur dispose, toutefois, de plusieurs logiciels. Thomas et Krebs (1997) ont présenté une évaluation de 29 logiciels, dont 13 spécialement conçus pour les analyses de puissance. Les capacités, la convivialité et les coûts de ces logiciels varient de façon importante. Ainsi, le chercheur qui désire investir dans un logiciel d'analyse de puissance complet et convivial devra débourser de 595 US $ pour *Power and precision* [*http://www.power-analysis.com/*] à 750 US $ pour *Nquery advisor* [*http://www.statsol.ie/html/nquery/nquery_home.html*] ou *Pass* [*http://www.ncss.com/pass.html*]. Cependant, le logiciel *G*Power* (Faul, Erdfelder, Lang et Buchner, 2007) en est à sa troisième version et demeure gratuit et disponible en ligne pour téléchargement [*http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/*]. Cette version, disponible pour les plates-formes *Windows XP* et *Vista* ainsi que *Mac-OS X 10.4*, remplace la version 2, configurée pour *MS DOS* et *Mac-OS 7-9*. Le calcul de puissance statistique *a priori* et *a posteriori* ainsi que le calcul de la taille de l'effet peuvent être effectués pour une variété de statistiques : khi-carré (qualité de l'ajustement ou test d'indépendance), *t* (différence de moyennes ou corrélation de Pearson), *F* (ANOVA à plan simple [*one-way*], univariée ou multivariée, régression linéaire) et *z*. Cette version ajoute trois autres types d'analyse de puissance : compromis entre α et β (le chercheur fixe le rapport β/α désiré), taille de l'effet minimum détecté et détermination de α en fonction de

la puissance statistique, de n et d'une taille d'effet donnée. *G*Power* se compare avantageusement à d'autres partagiciels, comme *Powpal* (Gorman, Primavera et Allison, 1995) ou *Power calculator* (Pittenger, 2001) et peut facilement combler les besoins des chercheurs ne désirant pas investir un montant significatif dans l'achat d'un logiciel commercial.

7. Conclusion

Nous avons vu que la recherche quantitative en sciences sociales s'en remet, presque uniquement, à l'hybridation entre les travaux de Fisher et ceux de Neyman et Pearson, qui caractérise actuellement les tests de signification (Loftus, 1996). Qui plus est, souvent, les études publiées manquent de puissance ou présentent des résultats statistiquement significatifs dont l'importance pratique est discutable (Vacha-Haase et Nilsson, 1998). De plus, la présence de variables confondues et la difficulté d'obtenir des échantillons aléatoires font de l'hypothèse nulle un *homme de paille* dont la réfutation n'a rien d'un exploit (Biskin, 1998 ; Loftus, 1996). L'omission d'interpréter la taille de l'effet contribuerait à élaborer une base de connaissances parfois contradictoires (Loftus, 1996 ; Maxwell, 2004), à entreprendre des recherches sans véritable chance de détecter un effet réel et à abandonner à tort des avenues prometteuses (Hallahan et Rosenthal, 1996). Il y a, il faut l'avouer, un problème éthique à entreprendre une recherche subventionnée par des fonds publics et dont le devis méthodologique équivaut, en termes de puissance, à jouer le résultat à pile ou face (dans leurs recherches, Cohen et ses successeurs ont systématiquement recensé des puissances moyennes de l'ordre de 50 %, soit une chance sur deux de déceler un effet statistiquement significatif).

Les causes du *statu quo* demeurent mal connues. Certains auteurs pointent du doigt la formation des chercheurs et les manuels de statistiques (Clark-Carter, 1997 ; Giguère, Hélie et Cousineau, 2004 ; Poitevineau, 2004) ; d'autres, la convivialité des logiciels statistiques (Blais, 1991 ; Foucart, 2001 ; Poitevineau, 2004), qui les rendrait plus accessibles aux profanes, alors que le sondage effectué par Mone, Mueller et Mauland (1996) révèle que les deux tiers des auteurs qui n'utilisent pas l'analyse de puissance affirment simplement que c'est parce que les périodiques dans lesquels ils publient ne l'exigent pas.

Comme nombre d'auteurs avant nous (American Psychological Association, 2001 ; Wilkinson and the Task Force on Statistical Inference, 1999), nous avons ensuite suggéré de suppléer les résultats des tests de signification avec l'interprétation de la taille de l'effet. Or, la taille de l'effet ne constitue pas une panacée aux problèmes méthodologiques en recherche (Sawilowsky, 2003) et n'est pas nécessairement comparable d'une étude à l'autre. Comme le fait justement remarquer Chow (1996), un résultat non significatif, malgré une taille d'effet considérable, ne permet pas automatiquement de conclure qu'un phénomène important est bien à l'œuvre dans la population. Cet état de faits peut plutôt suggérer que la taille de l'erreur-type associée aux estimateurs obtenus est grande. Williams, Zimmerman

et Zumbo (1995) ont d'ailleurs démontré qu'une augmentation de la fidélité d'une mesure, si elle est due à une diminution de la variance de l'erreur, entraîne une augmentation de la puissance. En d'autres mots, la décision du chercheur devrait être de revoir son devis méthodologique et ses instruments de mesure et de suspendre son jugement quant à la véracité de son hypothèse de recherche. Par ailleurs, d'autres problèmes affectent la rigueur des résultats de tests d'hypothèses : des devis méthodologiques qui ne respectent pas les conditions requises pour vérifier ou réfuter les hypothèses de recherche (Chow, 1996), la faible taille des échantillons et le manque de fiabilité des mesures, entre autres (Baguley, 2004 ; Fern et Monroe, 1996). De plus, des tests d'hypothèses sont appliqués malgré le non-respect plus ou moins important de leurs conditions d'application (échantillons aléatoires, normalité ou homogénéité de la variance, par exemple). Dans une telle situation, les valeurs obtenues pour *p*, la puissance statistique et la taille de l'effet seront biaisées (Kline, 2004). Conséquemment, il ne suffit pas d'indiquer une puissance statistique élevée et une taille d'effet importante ; encore faut-il que les postulats de base soient respectés, ou alors que le test utilisé soit suffisamment robuste.

Néanmoins, nous sommes d'avis que l'utilisation régulière de la taille de l'effet et de l'analyse de puissance *a priori* constituerait un progrès pour assurer la rigueur de la recherche quantitative en éducation.

Dans cette optique, nous encourageons les chercheurs désireux de publier leurs résultats de recherches quantitatives à adopter les pratiques qui suivent. D'abord, nous leur recommandons de vérifier si les conditions d'application des tests sont respectées et de publier les résultats de ces analyses préliminaires. Puis, lorsque des tests d'hypothèses sont effectués, il est préférable de publier les valeurs exactes obtenues pour *p* (plutôt que d'indiquer par des astérisques les résultats significatifs ou de donner uniquement le seuil de signification). Ensuite, nous suggérons fortement de préciser la taille des effets observés, surtout si la puissance statistique des tests effectués est élevée. Enfin, nous invitons les auteurs à exprimer leurs statistiques descriptives sous forme d'intervalles de confiance, au lieu de valeurs ponctuelles, afin que le lecteur ait une meilleure idée de la marge d'erreur associée à chaque statistique.

ENGLISH TITLE • Interpreting null-hypothesis significance tests : *p*, effect size, and power

SUMMARY • This paper aims to explain the rationale of hypothesis testing in research. Many studies have shown some confusion as to the significance of hypothesis testing results. This confusion could partially come from discrepancies between the approaches of Fisher, Neyman and Pearson as well as Bayes. The article clarifies the information provided by the significance coefficient, effect size and power, and proposes various software to perform power and effect size analysis.

KEY WORDS • hypothesis testing, statistical significance, effect size, power, inference.

TÍTULO EN ESPAÑOL • La interpretación de las pruebas de hipótesis: *p*, el tamaño del efecto y la potencia

RESUMEN • Este artículo tiene por objetivo explicitar la lógica de las pruebas de hipótesis en investigación. Varios estudios han demostrado una confusión en cuanto al significado de los resultados de las pruebas de hipótesis. Esta confusión tendría en parte por origen los puntos divergentes entre los enfoques de Fisher, de Neyman et Pearson así como de Bayes. El artículo precisa la información proporcionada por el coeficiente de significado, el tamaño del efecto y la potencia y propone distintos programas que permiten el análisis de la potencia y del tamaño del efecto.

PALABRAS CLAVES • **PRUEBAS DE HIPÓTESIS, SIGNIFICADO ESTADÍSTICO, TAMAÑO DEL EFECTO, POTENCIA, INFE**rencia.

Références

American Psychological Association (2001). *Publication manual of the American Psychological Association* (5e édition). Washington, District of Columbia: American Psychological Association.

Baguley, T. (2004). Understanding statistical power in the context of applied research. *Applied ergonomics, 35*(2), 73-80.

Behavioral and Brain Sciences, 21(2), 1998.

Bezeau, S. et Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of clinical and experimental neuropsychology, 23*(3), 399-406.

Biskin, B. H. (1998). Comment on significance testing. *Measurement and evaluation in counselling and development, 31*(1), 58-62.

Blais, J.-G. (1991). Statistique, méthodes quantitatives et analyse des données. *Repères, essais en éducation,* (13), 63-90.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard educational review, 48*(3), 378-399.

Cashen, L. H. et Geiger, S. W. (2004). Statistical power and the Testing of null hypotheses: a review of contemporary management research and recommendations for future studies. *Organizational research methods, 7*(2), 151-167.

Chow, S. L. (1996). *Statistical significance: rationale, validity and utility*. London, United Kingdom: Sage.

Clark-Carter, D. (1997). The account taken of statistical power in research. *British journal of psychology, 88*(1), 71-83.

Cohen, J. (1994). The Earth is round ($p < .05$). *American psychologist, 49*(12), 997-1003.

Cohen, J. (1988). *Statistical power analysis for the Behavioral sciences* (2e édition). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *Journal of abnormal and social psychology, 65*(3), 145-153.

Conseil des ministres de l'Éducation du Canada (2007). *Le Programme pancanadien d'évaluation (PPCE) et le Programme d'indicateurs du rendement scolaire (PIRS)* [En ligne]. Disponible le 25 juin 2007: *http://www.cmec.ca/pcap/indexf.stm*

Faul, F., Erdfelder, E., Lang, A.-G. et Buchner, A. (2007). G*Power 3 : a flexible statistical power analysis program for the social, behavioural, and biomedical sciences. *Behavior research methods*, *39*(2), 175-191.

Fern, E. F. et Monroe, K. B. (1996). Effect-size estimates : issues and problems in interpretation. *Journal of consumer research, 23*(2), 89-105.

Foucart, T. (2001). L'interprétation des résultats statistiques. *Mathématiques et sciences humaines, 39*(153), 21-28.

Gigerenzer, G. (1993). The Superego, the Ego, and the Id in statistical reasoning. Dans G. Keren et C. Lewis (Dir.) : *A handbook for data analysis in the Behavioral sciences : methodological issues*. Hillsdale, New Jersey : Erlbaum.

Giguère, G., Hélie, S. et Cousineau, D. (2004). Manifeste pour le retour des sciences en psychologie. *Revue québécoise de psychologie, 25*(3), 117-130.

Gorman, B. S., Primavera, L. H. et Allison, D. B. (1995). POWPAL : a program for estimating effect sizes, statistical power, and sample sizes. *Educational and psychological measurement, 55*(5), 773-776.

Hallahan, M. et Rosenthal, R. (1996). Statistical power : concepts, procedures, and applications. *Behaviour research and therapy, 34*(5/6), 489-499.

Jennions, M. D. et Møller, A. P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral ecology, 14*(3), 438-445.

Jones, L. V. et Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods, 5*(4), 411-414.

Kline, R. B. (2004). *Beyond Significance Testing : reforming data analysis methods in behavioral research.* Washington, District of Columbia : American Psychological Association.

Kosciulek, J. F. et Szymanski, E. M. (1993). Statistical power analysis of rehabilitation counselling research. *Rehabilitation counselling bulletin, 36*(4), 212-219.

Lecoutre, B., Poitevineau, J. et Lecoutre, M.-P. (2005). Une raison pour ne pas abandonner les tests de signification de l'hypothèse nulle. *Modulab, 33*, 243-248.

Lecoutre, M.-P. (1982). Comportements des chercheurs dans des situations conflictuelles d'analyse de données expérimentales. *Psychologie française, 27*(1), 1-8.

Lecoutre, M.-P. et Poitevineau, J. (2000). Aller au-delà des tests de signification usuels : vers de nouvelles normes de publication. *L'année psychologique, 100*(4), 683-713.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyse data. *Current Directions in Psychological Science, 5*(6), 161-171.

Maddock, J. E. et Rossi, J. S. (2001). Statistical power of articles published in three health psychology-related journals. *Health Psychology, 20*(1), 76-78.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research : causes, consequences, and remedies. *Psychological methods, 9*(2), 147-163.

Meehl, P. E. (1978). Theoretical risk and tabular asterisks : Sir Karl, Sir Ronald and the slow process of soft psychology. *Journal of consulting and clinical psychology, 46*(4), 806-834.

Michel, R., Ollivier-Gay, L., Spiegel, A. et Boutin, J.-P. (2002). Les tests statistiques : Intérêt, principe et interprétations. *Médecine tropicale, 62*(5), 561-563.

Millis, S. R. (2003). Statistical practices : the seven deadly sins. *Child Neuropsychology, 9*(3), 221-233.

Mone, M. A., Mueller, G. C. et Mauland, W. (1996). The perceptions and usage of Statistical Power in applied psychology and management research. *Personnel psychology, 49*(1), 103-120.

Morrison, D. E. et Henkel, R. E. (1970). *The significance test controversy.* Chicago, Illinois : Aldine.

Nakagawa, S. (2004). A Farewell to Bonferroni : the problems of Low Statistical Power and Publication Bias. *Behavioral ecology, 15*(6), 1044-1045.

Paul, K. M. et Plucker, J. A. (2004). Two steps forward, one step back : effect size reporting in gifted education research from 1995-2000. *Roeper review, 26*(2), 68-72.

Pittenger, D. J. (2001). Power calculator : a collection of interactive programs. *Educational and psychological measurement, 61*(2), 889-894.

Poitevineau, J. (2004). L'usage des tests statistiques par les chercheurs en psychologie : aspects normatif, descriptif et prescriptif. *Mathématiques et sciences humaines, 42*(3), 5-25.

Rosenthal, R., Rosnow, R. L. et Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research. A correlational approach.* Cambridge, United Kingdom : Cambridge University Press.

Rouanet, H. (1991). Les pratiques statisticiennes en question. Dans H. Rouanet, M.-P. Lecoutre, M.-C. Bert, B. Lecoutre, J.-M. Bernard et B. Leroux (Dir.) : *L'inférence statistique dans la démarche du chercheur.* Berne, Suisse : Peter Lang.

Sawilowsky, S. S. (2003). Deconstructing arguments from the case against Hypothesis Testing. *Journal of modern applied statistical methods, 2*(2), 467-474.

Shrout, P. E. (1997). Should significance tests be banned ? Introduction to a special section exploring the Pros and Cons. *Psychological science, 8*(1), 1-2.

Thomas, L. et Juanes, F. (1996). The importance of statistical power analysis : an example from animal behaviour. *Animal behaviour, 52*(4), 856-859.

Thomas, L. et Krebs, C. J. (1997). A review of statistical power analysis software. *Bulletin of the ecological society of America, 78*(2), 126-139.

Thompson, B. (1989). Statistical significance, result importance, and result generalizability : three noteworthy but somewhat different issues. *Measurement and evaluation in counselling and development, 22*(1), 2-6.

Vacha-Haase, T. et Nilsson, J. E. (1998). Statistical significance reporting : current trends and uses in *MECD*. *Measurement and evaluation in counselling and development, 31*(1), 46-57.

Vacha-Haase, T. et Thompson, B. (1998). Further comments on statistical significance tests. *Measurement and evaluation in counselling and development, 31*(1), 63-67.

Wilkinson, L. and the Task Force on Statistical Inference (1999). Statistical methods in psychology journals : guidelines and explanations. *American psychologist, 54*(8), 594-604.

Williams, R. H., Zimmerman, D. W. et Zumbo, B. D. (1995). Impact of measurement error on statistical power : review of an old paradox. *Journal of experimental education, 63*(4), 363-370.

Correspondance
jimmy.bourque@umoncton.ca
jean-guy.blais@umontreal.ca
Francois.Larose@USherbrooke.ca

Ce texte a été révisé par Sandra Najac.

Texte reçu le : 22 octobre 2007
Version finale reçue le : 25 août 2008
Accepté le : 25 septembre 2008